An Introduction to Phage Whole-Genome Sequencing and Annotation

Jason J. Gill

Department of Animal Science

Center for Phage Technology

Texas A&M University

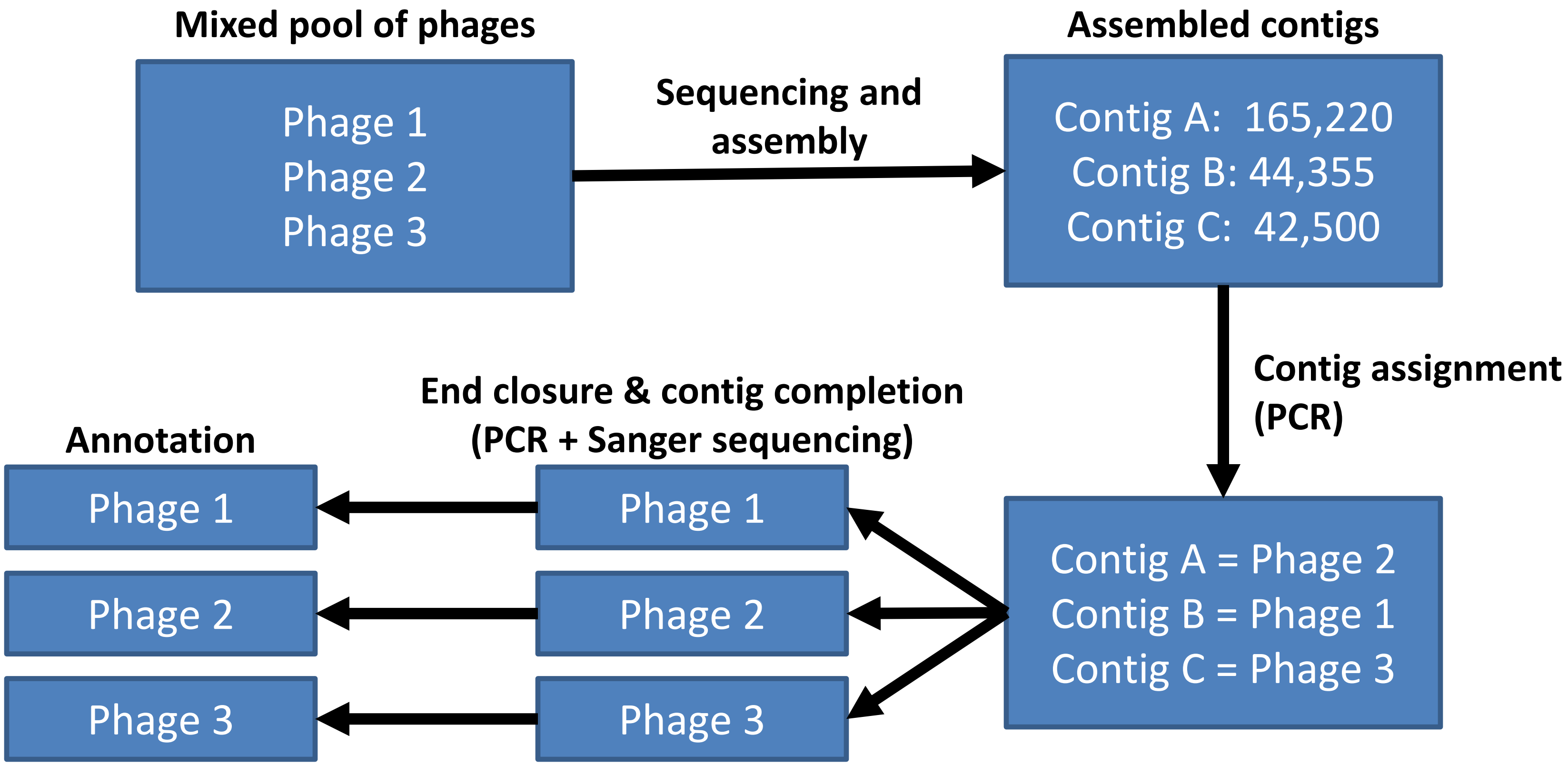
Center *for* Phage Technology

Texas A&M AgriLife Research

“Phage genomics”

- Not a narrowly-defined topic!
  - Whole-phage genome sequencing
  - Targeted phage metagenome sequencing
  - Metagenomics of viral consortia
  - Prophage mining/annotation

Whole phage genome sequencing and annotation



Mixed pool of phages

Phage 1
Phage 2
Phage 3

Sequencing and assembly

Assembled contigs

Contig A: 165,220
Contig B: 44,355
Contig C: 42,500

Contig assignment (PCR)

Contig A = Phage 2
Contig B = Phage 1
Contig C = Phage 3

End closure & contig completion (PCR + Sanger sequencing)

Phage 1
Phage 2
Phage 3

Annotation

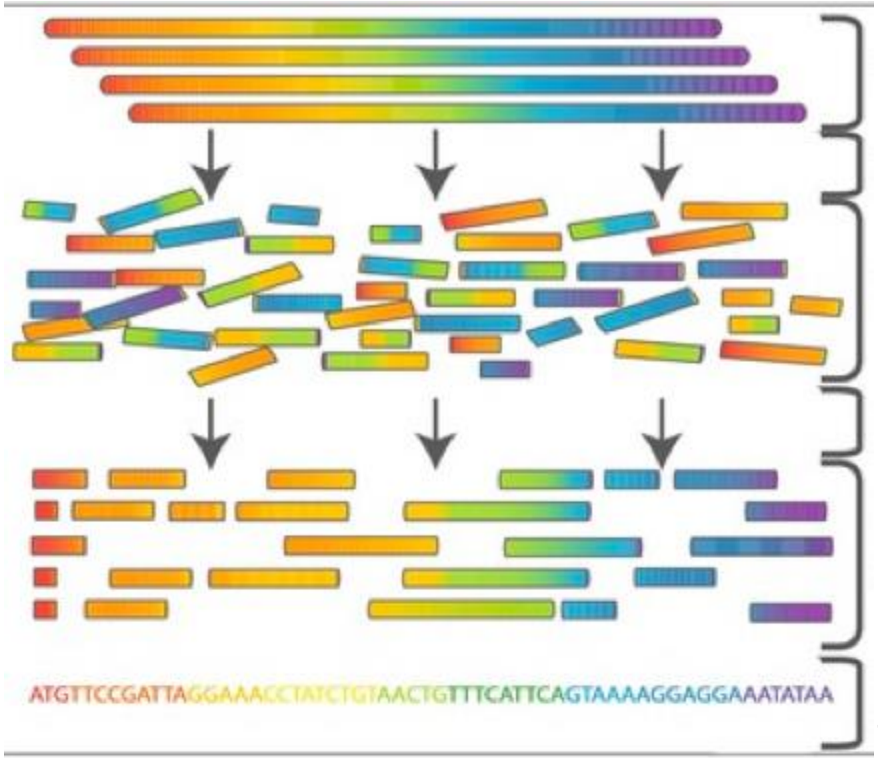
Phage 1
Phage 2
Phage 3

- Phages can usually be mixed into a single index or pool if they are not similar to each other
  - Different hosts
  - Different morphotypes

Center for Phage Technology

Texas A&M AgriLife Research

Shotgun sequencing



Genomic DNA

DNA is fragmented

Fragmented DNA is sequenced

Sequences are assembled

The assembly recapitulates the original order of the genome

The consensus sequence is produced from the assembly

Sequencing technology summary

Technology	Read length	Quality*	Total yield	Cost per base
Pyrosequencing (Ion Torrent)	400 – 600 bp	Moderate	Moderate	Moderate
Illumina	50 – 350 bp	High	High	Low
PacBio	2 – 20 kbp	Moderate	Moderate	Moderate
Nanopore	> 100 kbp ?	Low	Low-moderate	Moderate

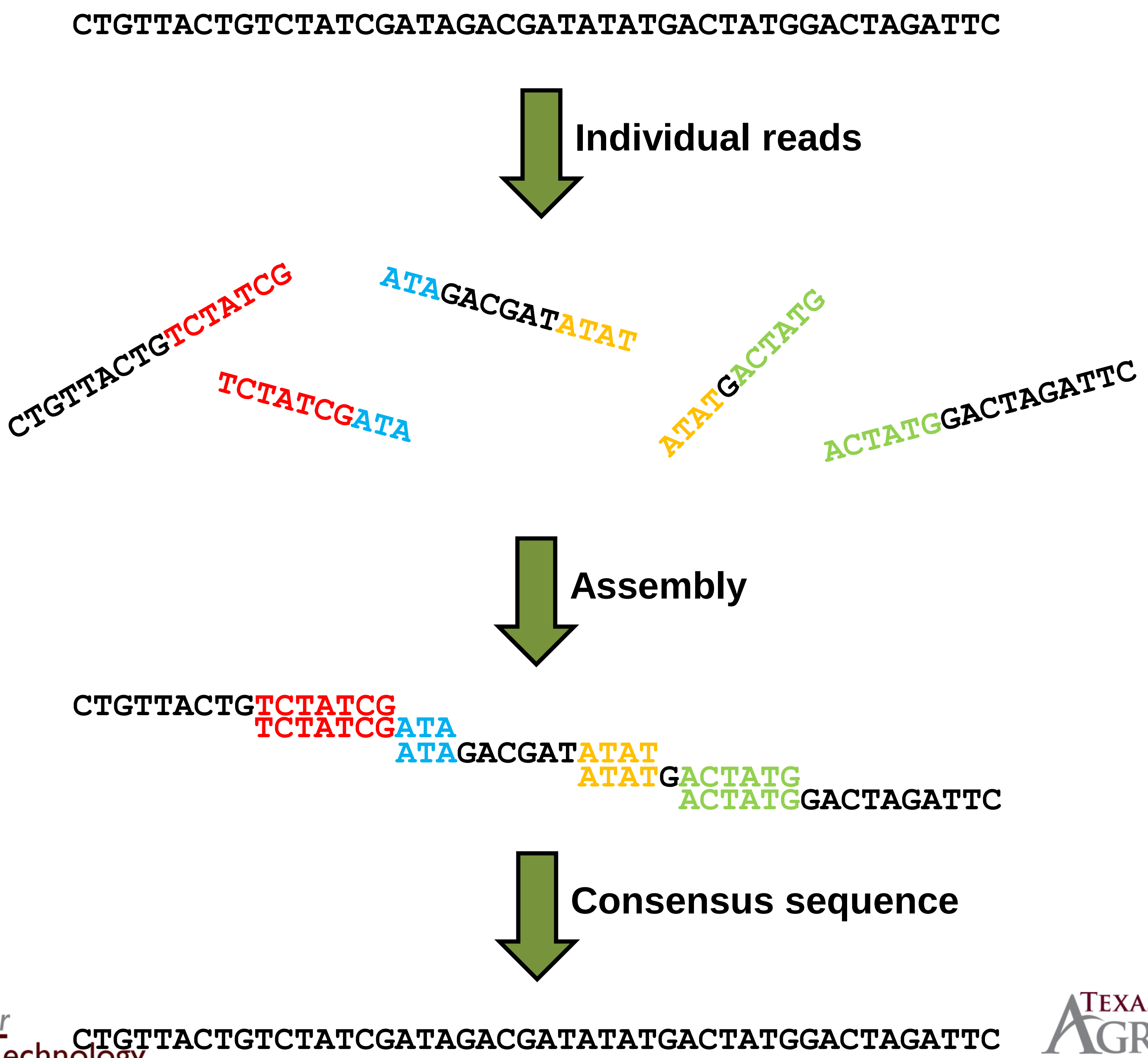
* Can vary as sequencing chemistries and software improve

Center for Phage Technology

Texas A&M AgriLife Research

Overlap-layout-consensus (OLC) assembly

- The "classic" method of assembly
  - Used for assembling long-read data (e.g., Sanger, PacBio and Oxford Nanopore reads)
- Reads can be of any length and can be non-uniform
- All sequence reads are compared pairwise to each other to find matches that meet a given threshold
  - $N(N-1)/2$ pairwise comparisons required for a set of N reads
- Higher tolerance of errors
- Assembly can be manually reviewed



CTGTTACTGTCTATCGATAGACGATATATGACTATGGACTAGATTC

Individual reads

CTGTTACTGTCTATCG

ATAGACGATATAT

TCTATCGATA

ATATGACTATG

ACTATGGACTAGATTC

Assembly

```
CTGTTACTGTCTATCG
         TCTATCGATA
               ATAGACGATATAT
                         ATATGACTATG
                              ACTATGGACTAGATTC
```

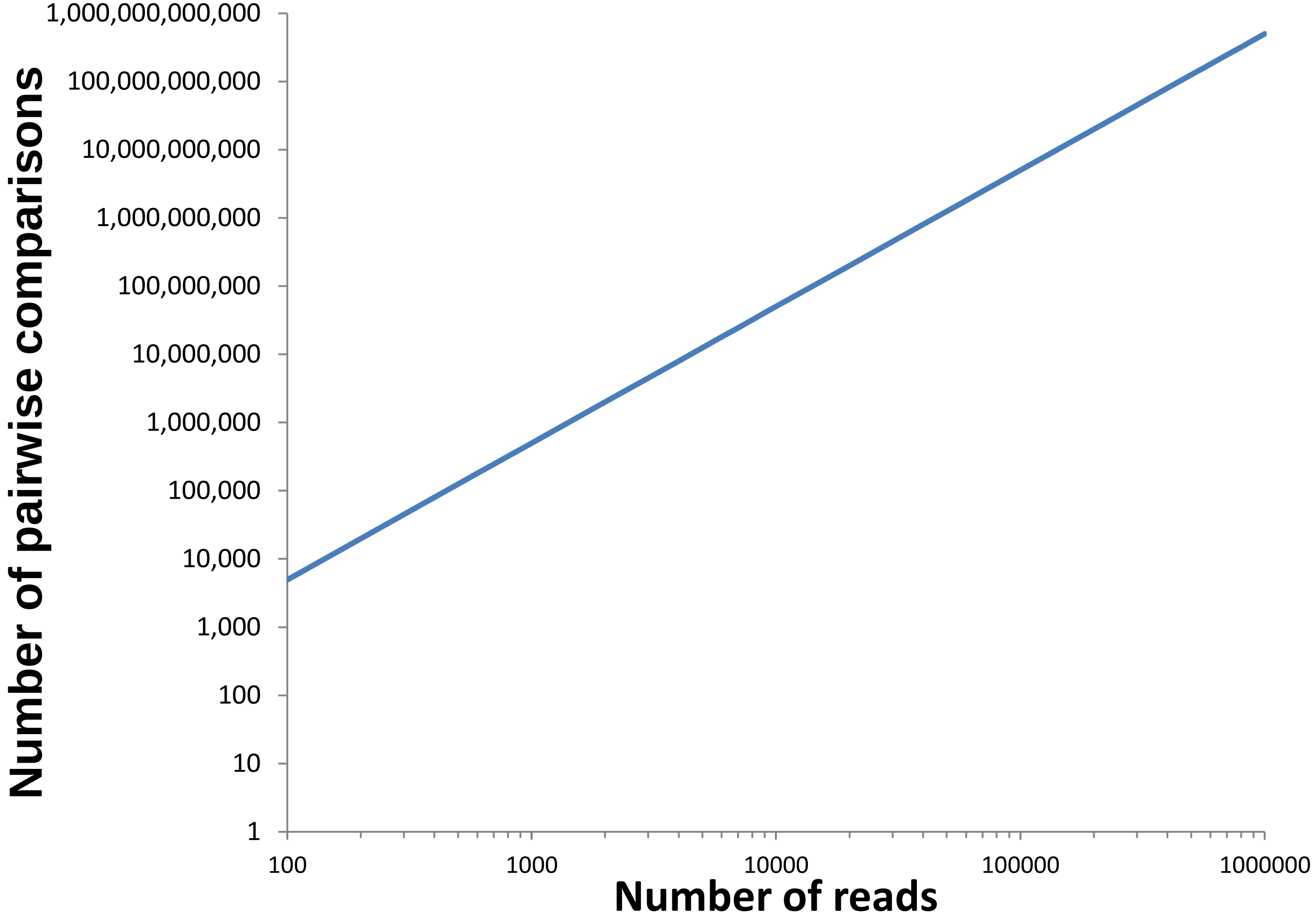
Consensus sequence

CTGTTACTGTCTATCGATAGACGATATATGACTATGGACTAGATTC

Center for Phage Technology

Texas A&M AgriLife Research

Overlap consensus: number of reads vs. number of pairwise comparisons



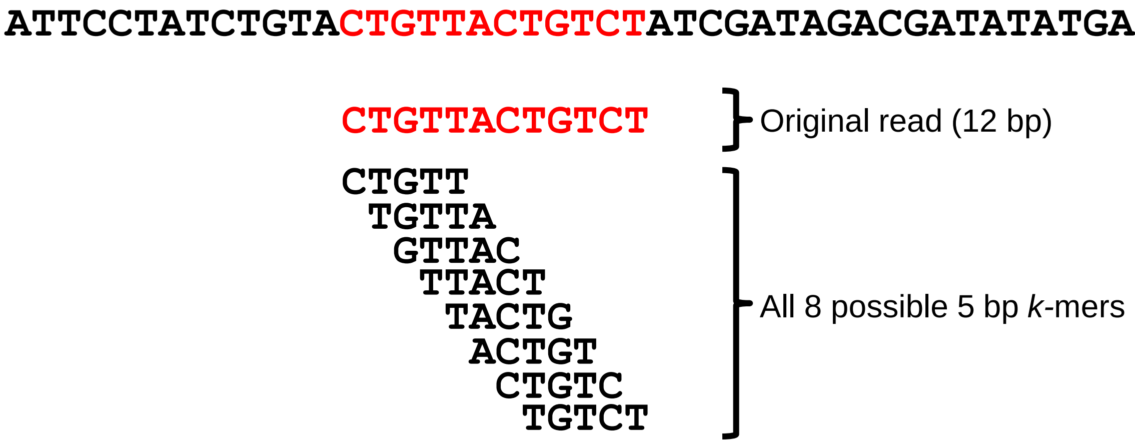
Center for Phage Technology

Texas A&M AgriLife Research

de Bruijn graph assembly

- All reads are split into sequences of a defined length, called a *k*-mer
- Identical *k*-mers are collapsed into a single *k*-mer, reducing computational requirements
    - Redundant *k*-mers are discarded
    - All remaining *k*-mers are unique
- *k*-mers can only be linked in the assembly if they are *identical* and offset by *one position*
- The entire genome can only be assembled if this chain of single-offset *k*-mers is unbroken
- Assumptions for complete assembly:
    - All *k*-mers in the genome are contained in the read set
    - All *k*-mers are error-free
    - Each *k*-mer appears only once in the genome

k-mer generation



```
ATTCCTATCTGTACTGTTACTGTCTATCGATAGACGATATATGACTATGGACTAGATTC

             CTGTTACTGTCT       Original read (12 bp)

             CTGTT
              TGTTA
               GTTAC
                TTACT
                 TACTG          All 8 possible 5 bp k-mers
                  ACTGT
                   CTGTC
                    TGTCT
```

- In practice, *k*-mers of 21 to >100 are used for assembly of phage genomes

Assembly algorithms overview

- **Overlap-layout-consensus (OLC)**
  - Searches for overlaps in all-against-all pairwise comparisons
  - Computationally more intensive
  - More tolerant of low quality data
  - More suited to long-read, low-coverage assemblies
  - A more intuitive process

- **De Bruijn graph (DBG)**
  - Splits reads into *k*-mers and assembles based on De Bruijn graphs (links overlapping k-mers shifted by one position at a time)
  - Computationally more efficient at high coverage depth (identical *k*-mers are merged)
  - Less tolerant of low quality data (errors force *k*-mers to remain separate)
  - Better for short-read, high-coverage assemblies

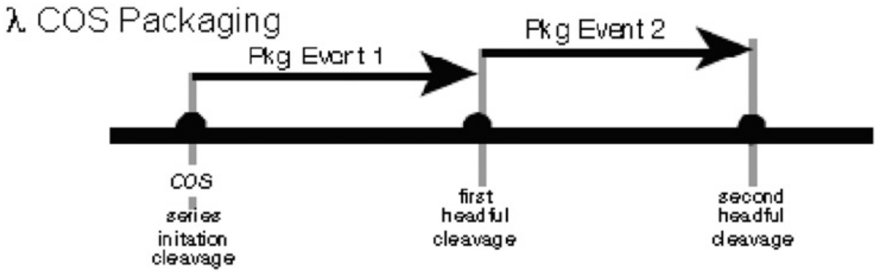
Assembly programs

- **Overlap-layout-consensus (OLC)**
  - **Phrap**
  - **Celera**
  - **Newbler (454)**
  - Phusion
  - **Allora (PacBio)**
  - Sequencher

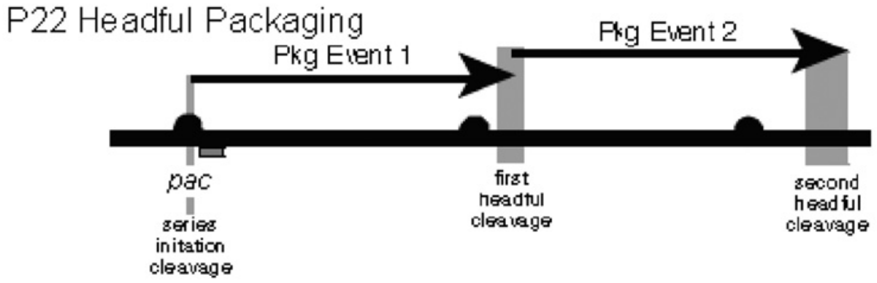
- **De Bruijn graph (DBG)**
  - Euler
  - ABySS
  - **Velvet**
  - **SOAPdenovo**
  - **SPAdes**
  - CLC bio Genomics

Phage DNA packaging strategies

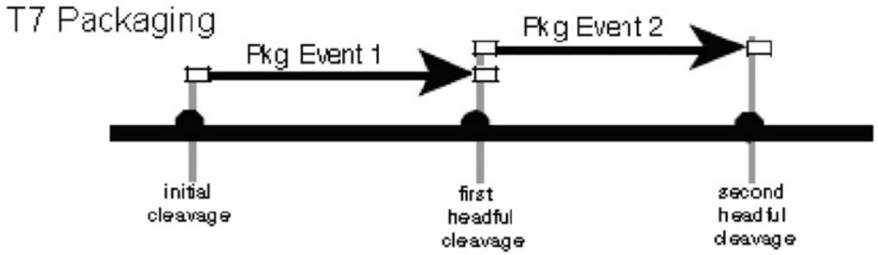
Determining your phage termini



- gDNA with short 5’ or 3’ overhangs
- Assembly **may** or **may not** have *cos* termini at the end of the contig
- Genome should be opened at *cos* ends



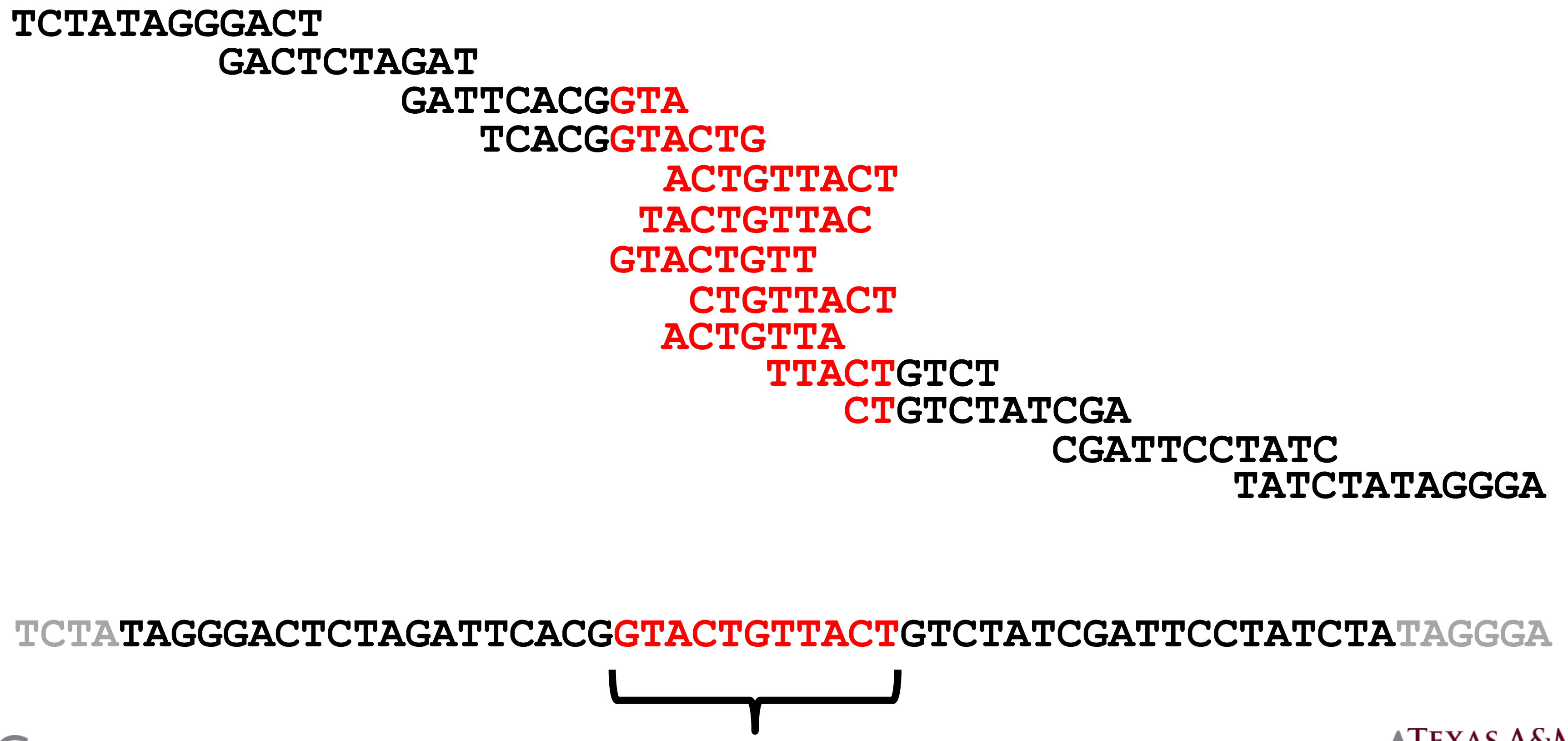
- Genome has no fixed or “true” termini
- Genome is reopened to convention (e.g., between *rIIAB* for T4-like phages)



- Terminal repeats are collapsed in the middle of the contig
- Must be determined bioinformatically or experimentally

Assembly of repeat regions

GTACTGTTACTGTCTATCGATTCCTATCTATAGGGACTCTAGATTCACGGTACTGTTACT



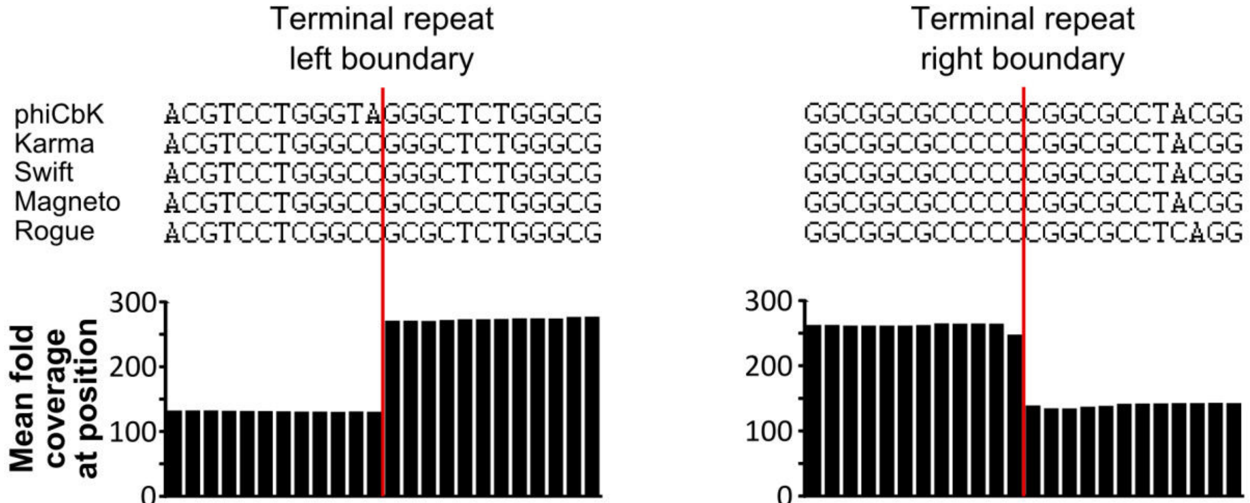
```
TCTATAGGGACT
        GACTCTAGAT
               GATTCACGGTA
                  TCACGGTACTG
                         ACTGTTACT
                        TACTGTTAC
                       GTACTGTT
                          CTGTTACT
                         ACTGTTA
                             TTACTGTCT
                                CTGTCTATCGA
                                        CGATTCCTATC
                                               TATCTATAGGGA

TCTATAGGGACTCTAGATTCACGGTACTGTTACTGTCTATCGATTCCTATCTATAGGGA
```

High coverage region

Terminal repeat boundaries

- Terminal repeats in phage genomes like T7 or T5 may be detectable by analyzing sequence coverage



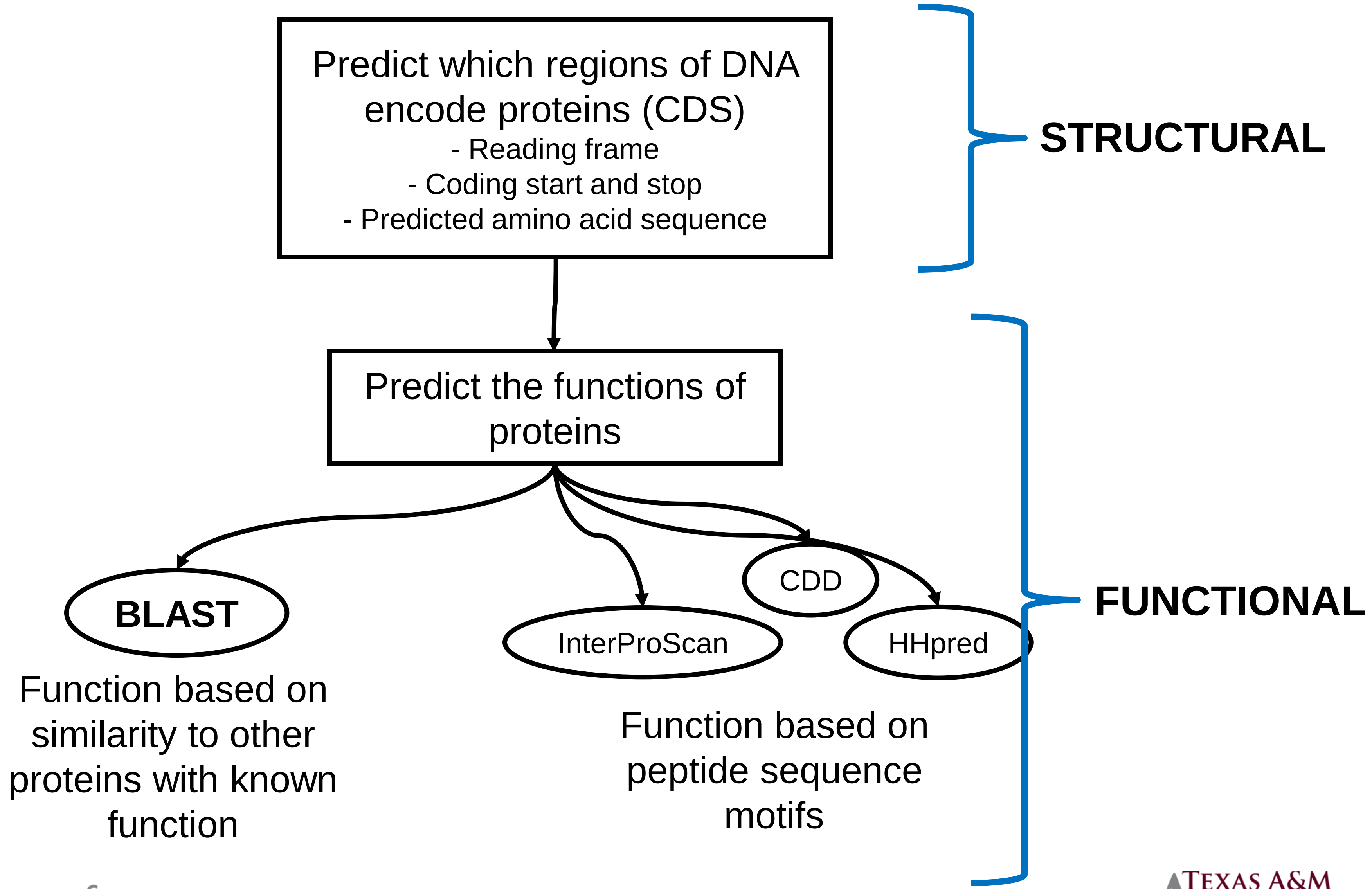
Gill *et al*. BMC Genomics 2012

- PhageTerm is available to automate this analysis and find genomic termini
  - – http://www.biorxiv.org/content/early/2017/02/15/108100

Genome annotation workflows

- Environmental / Metagenomic
  - – Identification of genes/proteins/pathways from metagenomic assembly
  - – Individual phages often not cultured
  - – Often emphasis on relationships, distribution, ecology
- Whole genome
  - – Annotation of individual complete, closed genomes
  - – Often emphasis on presence of toxins/virulence determinants, determination of phage lifestyle
  - – Basis for taxonomy, future genetic or molecular biology experiments

Genome annotation



Predict which regions of DNA encode proteins (CDS)

- Reading frame
- Coding start and stop
- Predicted amino acid sequence

STRUCTURAL

Predict the functions of proteins

FUNCTIONAL

BLAST

Function based on similarity to other proteins with known function

InterProScan

CDD

HHpred

Function based on peptide sequence motifs

Center for Phage Technology

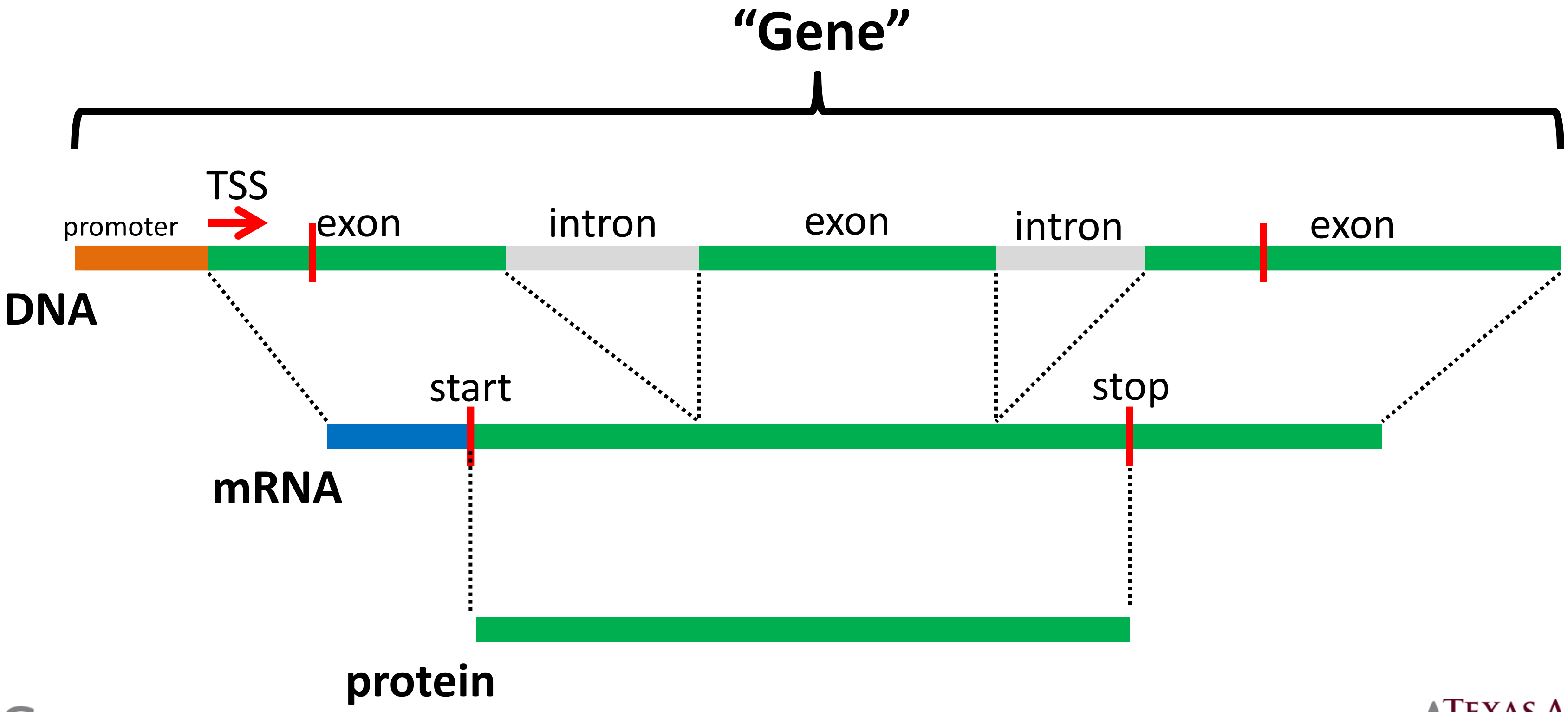
Texas A&M AgriLife Research

Structural annotation tools

- For protein-coding genes
  - GeneMark
    - http://exon.gatech.edu/GeneMark/
  - MetaGeneAnnotator
    - http://metagene.cb.k.u-tokyo.ac.jp/
  - Glimmer3
    - http://ccb.jhu.edu/software/glimmer/index.shtml
  - Prodigal
    - http://prodigal.ornl.gov/server.html
- For non-coding features
  - tRNAScan
    - http://lowelab.ucsc.edu/tRNAscan-SE/
  - ARAGORN
    - http://mbio-serv2.mbioekol.lu.se/ARAGORN/
  - TransTermHP
    - http://transterm.cbcb.umd.edu/

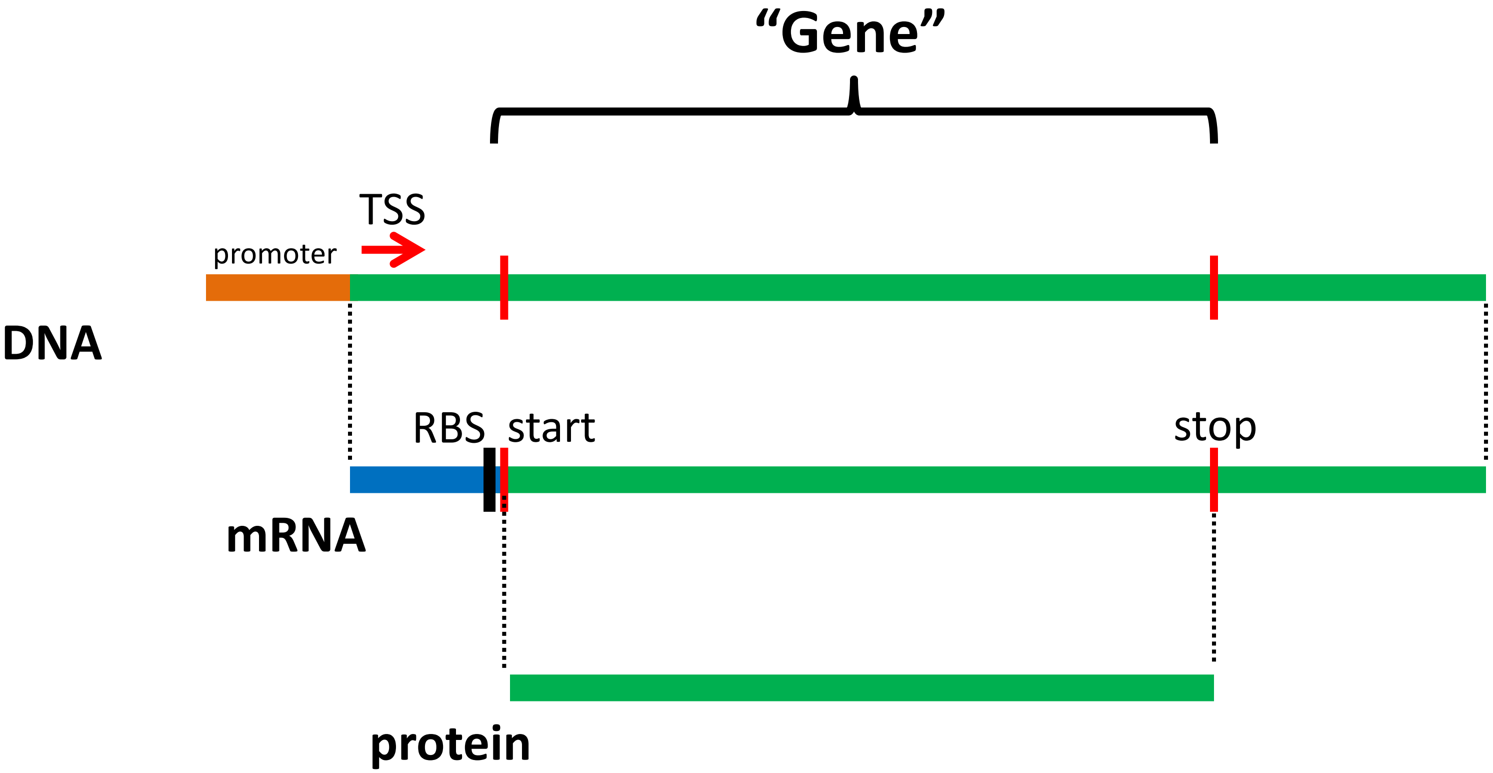
Eukaryotic gene

- Extensive mRNA processing for intron splicing, 5’ and 3’ modification
- Difficult to infer protein sequence directly from DNA sequence



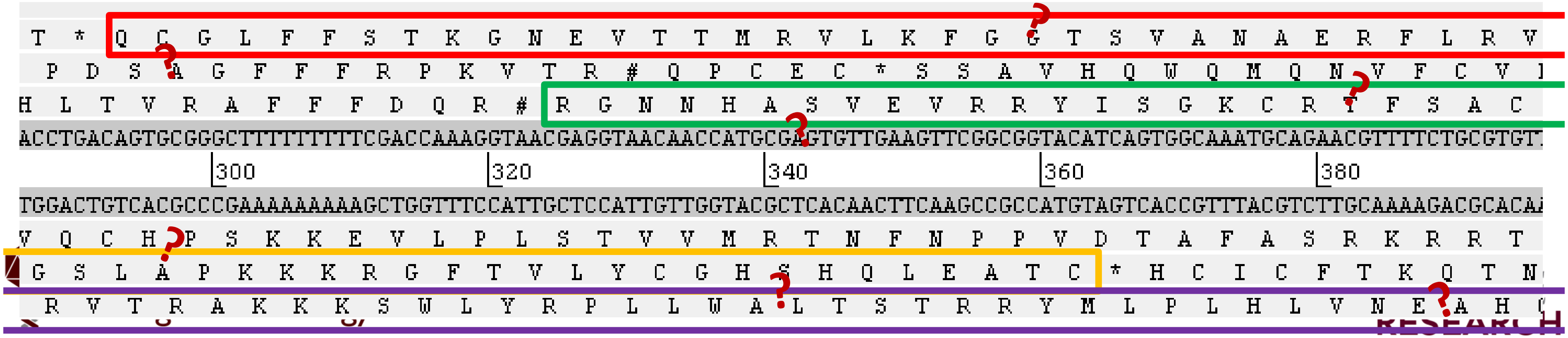
Prokaryotic gene

- Introns rare, little mRNA processing
- Easy to infer protein sequence directly from DNA sequence



Protein-coding genes

- DNA consists of two complementary strands with three possible reading frames for each, **six reading frames total**
- Just by looking at the DNA sequence, it is not obvious which strand and reading frame encodes a protein
- A gene must be an ORF, but not any ORF can be a gene!



```
 T  *  Q  C  G  L  F  F  S  T  K  G  N  E  V  T  T  M  R  V  L  K  F  G  G  T  S  V  A  N  A  E  R  F  L  R  V
  P  D  S  A  G  F  F  F  R  P  K  V  T  R  #  Q  P  C  E  C  *  S  S  A  V  H  Q  W  Q  M  Q  N  V  F  C  V
H  L  T  V  R  A  F  F  F  D  Q  R  #  R  G  N  N  H  A  S  V  E  V  R  R  Y  I  S  G  K  C  R  T  F  S  A  C
ACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGT
              |300                |320                |340                |360                |380
TGGACTGTCACGCCCGAAAAAAAAAGCTGGTTTCCATTGCTCCATTGTTGGTACGCTCACAACTTCAAGCCGCCATGTAGTCACCGTTTACGTCTTGCAAAAGACGCACAA
V  Q  C  H  P  S  K  K  E  V  L  P  L  S  T  V  V  M  R  T  N  F  N  P  P  V  D  T  A  F  A  S  R  K  R  R  T
 G  S  L  A  P  K  K  K  R  G  F  T  V  L  Y  C  G  H  S  H  Q  L  E  A  T  C  *  H  C  I  C  F  T  K  Q  T  N
  R  V  T  R  A  K  K  K  S  W  L  Y  R  P  L  L  W  A  L  T  S  T  R  R  Y  M  L  P  L  H  L  V  N  E  A  H
```

Gene prediction tools

- In addition to a start, RBS and stop codon, a gene encodes biologically meaningful information, which means its DNA sequence is **not random**
- Gene prediction tools use this fact to locate probable coding sequences
  - GC content of the 3rd codon position reflects genomic GC content
  - Frequency and distribution of dinucleotide pairs
  - Periodicity of Fourier-transformed DNA sequence
  - Hidden Markov Models
  - Codon usage bias compared to organism as a whole
- We use two programs for predicting phage genes, **MetaGeneAnnotator (MGA)** and **Glimmer3**
- These tools are generally accurate (> 90%) but still need some manual curation of the output

Gene prediction and translation initiation

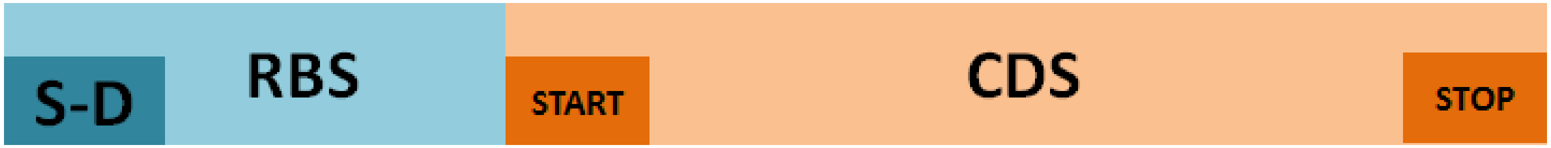
- Presumably, all protein-coding genes must be translated into protein from an mRNA, which requires **initiation**
- A Translation Initiation Site (TIR) consists of a **Shine-Dalgarno (S-D)** sequence, a 4-12 bp spacer, and a start codon
  - The S-D sequence must base-pair with the complementary sequence at the 3' end of the 16S rRNA to initiate translation of a protein
- The **strength** of translation initiation is affected by how close a gene's RBS is to the consensus S-D sequence **AGGAGGT**
- **Any 3-base subset** of the canonical S-D can be used in a TIR
  - Must have appropriate spacing
  - Wobble base-pairing rules apply

Valid Shine-Dalgarno sequences

Watson-Crick	Wobble (G-U)
AGGAGGT	AGGAGGT
AGGAGG	GGGGGG
GGAGGT	GGGGGT
AGGAG	AGGGG
GGAGG	GGGGG
GAGGT	GGGGT
AGGA	GGGA
GGAG	GGGG
GAGG	GGGT
AGGT	GGG
AGG	
GGA	
GAG	
GGT	

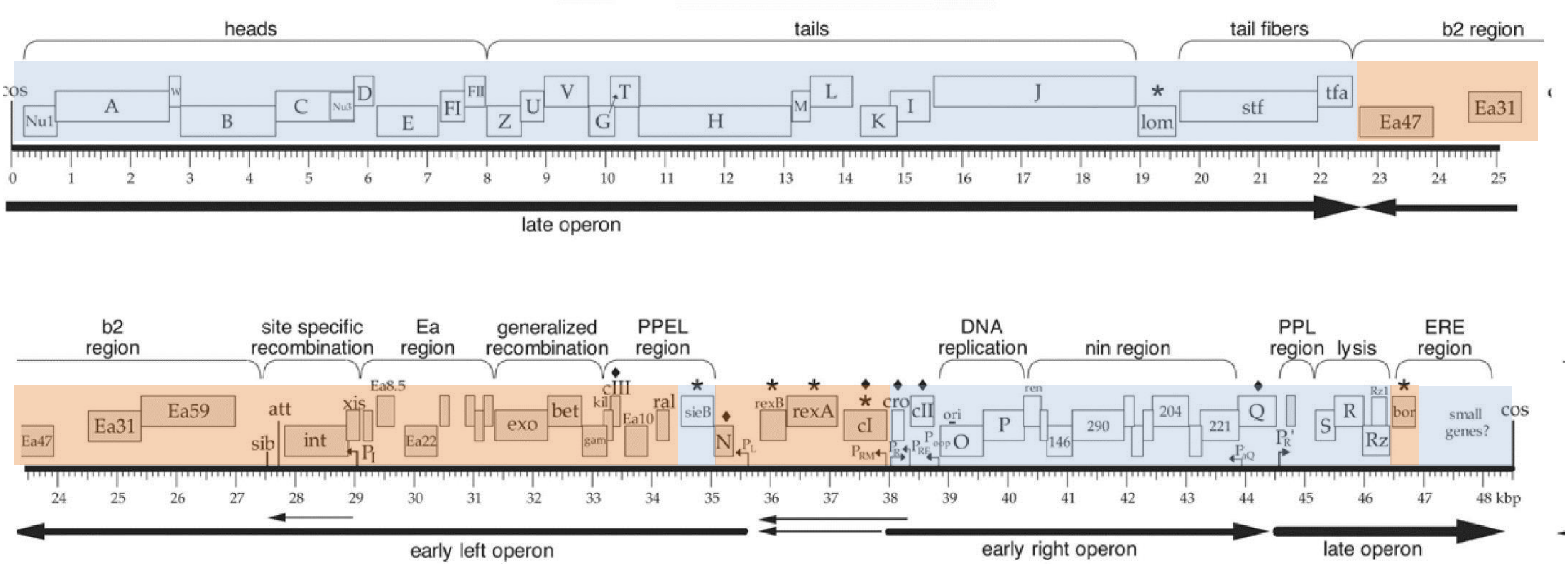
Basic gene structure

- A protein-coding gene must:
  - Have a translational start signal upstream of a valid start codon (ATG, GTG, TTG)
  - Encode a protein in an **open reading frame (ORF)** determined by the start codon (also called the **coding segment**, or CDS)
  - Be terminated by a stop codon



```
AGTAGGTACCTGATTATGCAGCATGTG...TCGGATTAAGCTT
               M  R  H  V      S  D  *
```

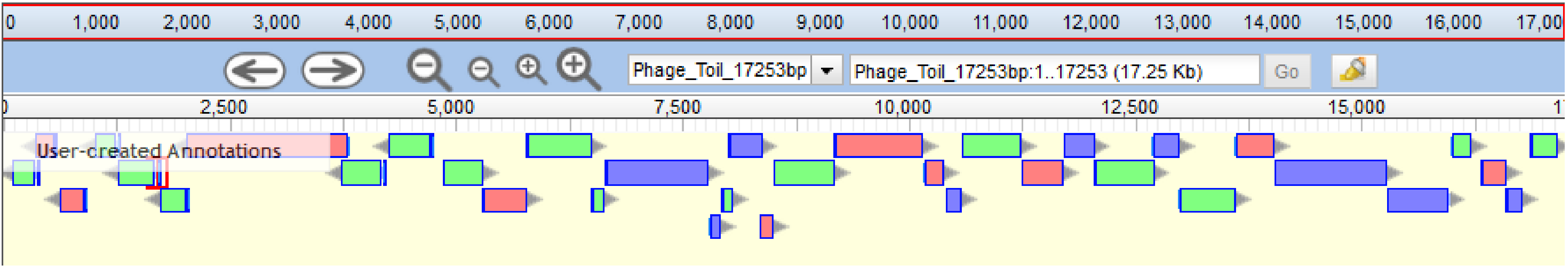
Coding density and organization



- **Density**
  - Most phages have coding densities of >90%
  - Most of the DNA contains some kind of feature: protein coding gene, tRNA, terminator, regulatory element, etc.
  - These features are **tightly packed** and may even **overlap** if biologically possible
- **Transcriptional units**
  - Phage genes are translated from **polycistronic mRNA's**
  - Genes tend to be arranged in groups on the plus or minus strand

General gene finding rules for phage

- Phages have high coding density
  - Genes tend to have minimal gaps between them or overlap slightly (up to ~5-8 aa)
  - Genes should never be embedded in each other on opposite strands
- Genes tend to be arranged into transcriptional units: blocks of genes on one strand or the other
- Start codons: ATG > GTG >> TTG
- Have recognizable **translation initiation sites**, but only a few will have the full consensus S-D sequence AGGAGGT
- Most genes will encode proteins > 30 aa
- Sometimes there is no good-looking gene for a given DNA region and **that is OK**
  - There may be a regulatory element or some other function for that sequence



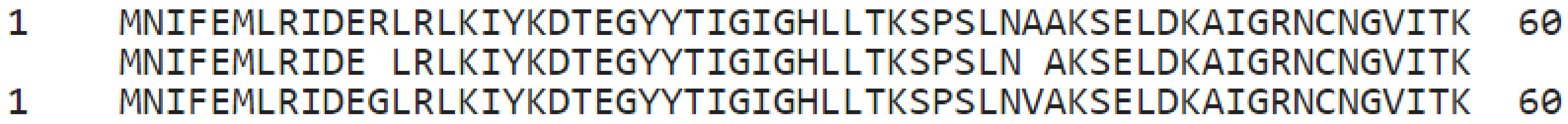
BLAST: Basic Local Alignment Search Tool

- The **database** used will determine the scope of your search
- There are **many** databases available to be searched by BLAST
  - **nr** (non-redundant database): the default at NCBI, contains all unique deposited sequences
  - **SwissProt**: Manually curated protein dataset from EMBL
  - **TrEMBL**: Electronically inferred annotations from SwissProt
  - **UniRef**: Clusters of homologous proteins in UniProt
  - FigFams, COGs, POGs, ARDB, mVirDB, etc.

BLAST of T4 E vs. the nr database

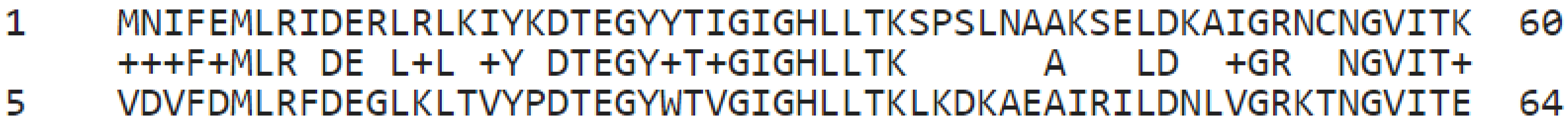
- In theory an E value < 1 is significant
- In practice, E values of < 1e-3 or 1e-5 are considered relevant, **if they cover most or all of the protein**

T4 vs. RB14, E = 9e-115



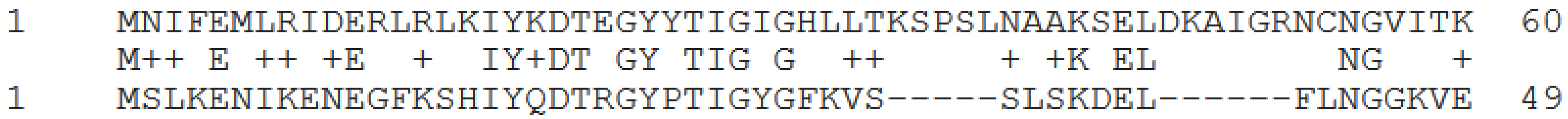
```
1     MNIFEMLRIDERLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITK  60
      MNIFEMLRIDE LRLKIYKDTEGYYTIGIGHLLTKSPSLN AKSELDKAIGRNCNGVITK
1     MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNVAKSELDKAIGRNCNGVITK  60
```

T4 vs. Phi92, E = 8e-52



```
1     MNIFEMLRIDERLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITK  60
      +++F+MLR DE L+L +Y DTEGY+T+GIGHLLTK      A   LD  +GR  NGVIT+
5     VDVFDMLRFDEGLKLTVYPDTEGYWTVGIGHLLTKLKDKAEAIRILDNLVGRKTNGVITE  64
```

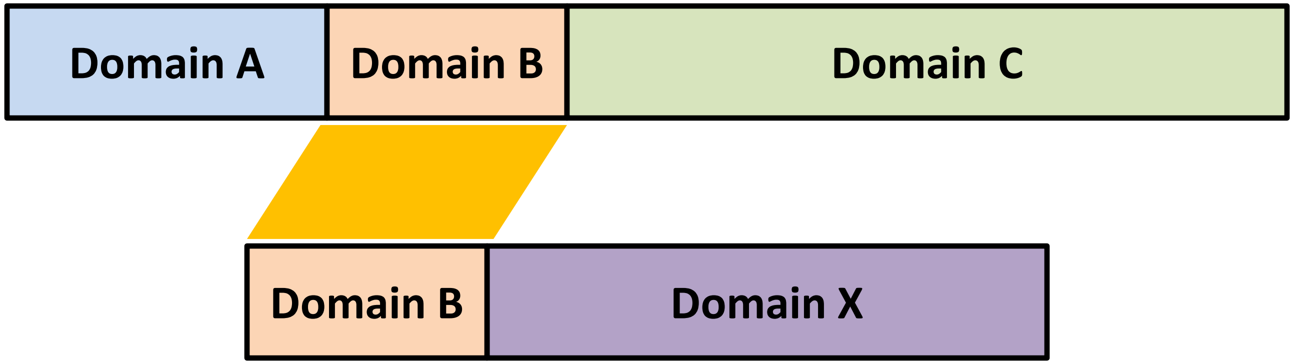
T4 vs. *C. concisus*, E = 7e-08



```
1     MNIFEMLRIDERLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITK  60
      M++ E ++ +E  +  IY+DT GY TIG G  ++     + +K EL        NG   +
1     MSLKENIKENEGFKSHIYQDTRGYPTIGYGFKVS-----SLSKDEL------FLNGGKVE  49
```

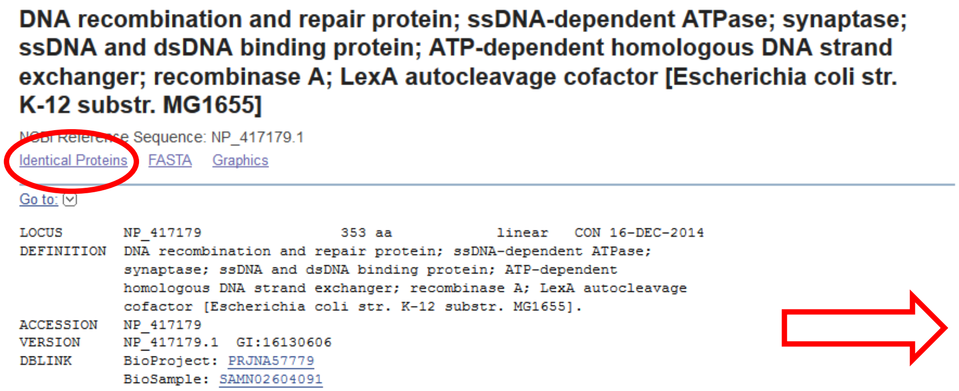
Partial protein similarity can lead to misleading results

- Two different proteins can share a region of similarity if they share a functional domain
- E.g., both proteins may hydrolyze ATP but otherwise have different functions
- BLAST E-value can be misleading if a there is a good match over part of a protein



WP numbers

- To save database space and improve speed, identical protein sequences are now collapsed into a single record with a **WP_** accession number
- A single representative record is chosen to be the “face” of the group
- The record chosen is not necessarily the most informative, and may not be the one you’re looking for!



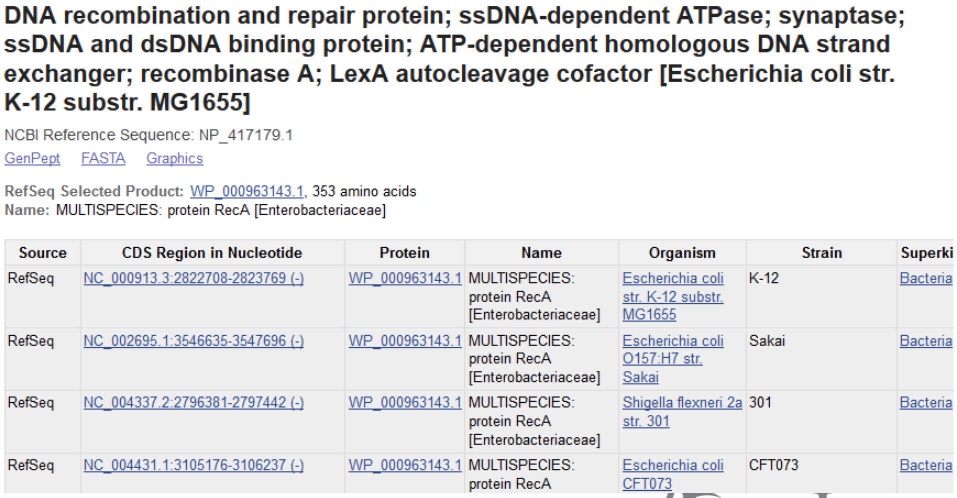
DNA recombination and repair protein; ssDNA-dependent ATPase; synaptase; ssDNA and dsDNA binding protein; ATP-dependent homologous DNA strand exchanger; recombinase A; LexA autocleavage cofactor [Escherichia coli str. K-12 substr. MG1655]

NCBI Reference Sequence: NP_417179.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_417179                353 aa            linear   CON 16-DEC-2014
DEFINITION  DNA recombination and repair protein; ssDNA-dependent ATPase;
            synaptase; ssDNA and dsDNA binding protein; ATP-dependent
            homologous DNA strand exchanger; recombinase A; LexA autocleavage
            cofactor [Escherichia coli str. K-12 substr. MG1655].
ACCESSION   NP_417179
VERSION     NP_417179.1  GI:16130606
DBLINK      BioProject: PRJNA57779
            BioSample: SAMN02604091
```



DNA recombination and repair protein; ssDNA-dependent ATPase; synaptase; ssDNA and dsDNA binding protein; ATP-dependent homologous DNA strand exchanger; recombinase A; LexA autocleavage cofactor [Escherichia coli str. K-12 substr. MG1655]

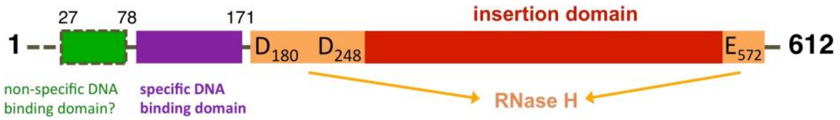
NCBI Reference Sequence: NP_417179.1

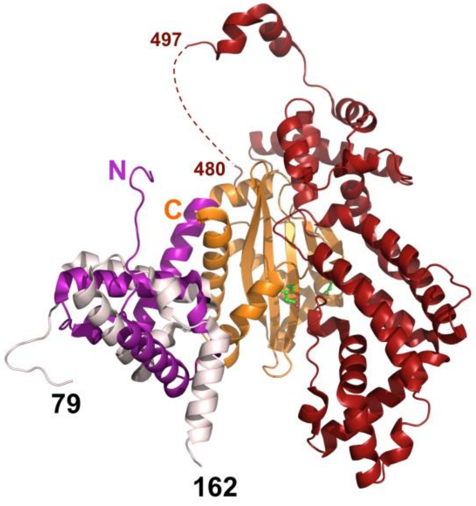
GenPept FASTA Graphics

RefSeq Selected Product: WP_000963143.1, 353 amino acids
Name: MULTISPECIES: protein RecA [Enterobacteriaceae]

Source	CDS Region in Nucleotide	Protein	Name	Organism	Strain	Superki
RefSeq	NC_000913.3:2822708-2823769 (-)	WP_000963143.1	MULTISPECIES: protein RecA [Enterobacteriaceae]	Escherichia coli str. K-12 substr. MG1655	K-12	Bacteria
RefSeq	NC_002695.1:3546635-3547696 (-)	WP_000963143.1	MULTISPECIES: protein RecA [Enterobacteriaceae]	Escherichia coli O157:H7 str. Sakai	Sakai	Bacteria
RefSeq	NC_004337.2:2796381-2797442 (-)	WP_000963143.1	MULTISPECIES: protein RecA [Enterobacteriaceae]	Shigella flexneri 2a str. 301	301	Bacteria
RefSeq	NC_004431.1:3105176-3106237 (-)	WP_000963143.1	MULTISPECIES: protein RecA	Escherichia coli CFT073	CFT073	Bacteria

Conserved domain searches





E. coli RNase H

1RNH: Yang et al., Science (1990) **249**:1398- 1405.

- Many proteins are organized into **functional domains,** each of which contributes to the protein's function
  - Ligand binding domains
  - Enzymatic active sites
  - Cofactor binding sites
  - Structural components
  - Etc.
- Some have argued that the **domain** is the smallest meaningful biological unit, rather than the gene
- Domains can be reshuffled to form proteins with new functions

Conserved domain searches

- Exact methods vary, but these tools search your query sequence against **models of functional domains** rather than individual sequences as in BLAST
  - **NCBI Conserved Domain Database (CD-Search):** Includes NCBI data and 5 external databases
    - Fast, allows batch searches online
  - **EMBL InterProScan:** Integrates 14 member databases into a unified system of functional domains
    - Slow, online search allows 1 sequence at a time
  - **HHpred (Tuebingen MPI):** Very sensitive dynamic searches of models against models
    - Slow, 1 sequence at a time, output can be difficult to interpret

Genome annotation tools

Fully automated annotation

- RAST/myRAST
  - http://rast.nmpdr.org/
- Prokka
  - http://www.vicbioinformatics.com/software.prokka.shtml
- NCBI Prokaryotic Pipeline
  - https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

Semi-automated annotation

- DNA Master
  - http://cobamide2.bio.pitt.edu/
- CPT Galaxy/Apollo
  - https://cpt.tamu.edu/galaxy-pub/

Manual annotation / genome editors

- Sanger Artemis
  - http://www.sanger.ac.uk/science/tools/artemis
- Broad Argo
  - https://archive.broadinstitute.org/annotation/argo/

Center for Phage Technology

Texas A&M AgriLife Research